Artificial intelligence literacy development in preservice teachers through metacognitive reflection a mixed methods study

Gabriele Biagini[*], Maria Ranieri[^]

Abstract
This study examines how future teachers' AI literacy develops throughout a four-module professional development program and how metacognitive reflection supports transformative learning. Sixty-three Italian teacher-learners produced short reflective diaries after each module. Using a convergent mixed-methods design, we integrated computational text analyses with qualitative coding to characterize change across cognitive, operational, critical, and ethical dimensions of AI literacy. Texts were pre-processed in Italian with spaCy (lemmatization and extended stopword filtering) and represented with TF-IDF to compute inter-section cosine similarity between phases and within-participant coherence; lexical diversity (type-token ratio), unsupervised topic modeling (LDA, k = 5), and sentiment scores complemented the analysis. Qualitative thematic analysis contextualized patterns with coder agreement and representative excerpts. Results indicate declining inter-diary similarity across modules alongside stable or rising within-participant coherence, suggesting cohort-level conceptual reorganization while individuals consolidate their understanding. Lexical diversity decreases modestly as vocabulary specializes; topic distributions shift from generic to practice-oriented themes; and sentiment

[*] University of Florence, Florence, Italy. Corresponding author, e-mail:gabriele.biagini@unifi.it.
[^] University of Florence, Florence, Italy. E-mail: maria.ranieri@unifi.it.
The submission is the product of the joined effort of all authors. The shared aspects of the work, including conceptualisation, methodology, and manuscript development, were carried out jointly by all authors. It is also possible to attribute the following contributions to Gabriele Biagini: Software, Formal Analysis, Investigation, Data Curation, Writing – Original Draft, Visualisation; and to Maria Ranieri: Validation, Resources, Writing – Review & Editing, Supervision, Project Administration.
We wish to express our sincere gratitude to the pre-service teachers who participated in this study. Their thoughtful engagement and deep reflections were invaluable and formed the foundation of this research. We would also like to thank our colleagues at the University of Florence for their feedback during the preliminary stages of this work.

Excellence and Innovation in Learning and Teaching (ISSNe 2499-507X), 2026, 1

Doi: 10.3280/exioa1-2026oa23214

evolves from uncertainty to more balanced, professionally grounded appraisals. Taken together, the findings portray a progression from initial ambiguity toward increasingly articulated AI literacy, linking computational signals with reflective evidence of transformation. We discuss implications for teacher-education curricula that aim to build sustainable, ethically aware AI capacities without presupposing technical backgrounds.
Keywords*:* artificial intelligence, teacher education, transformative learning, metacognition, AI literacy, mixed methods, natural language processing.

Article submitted:16/04/2026; accepted: 24/06/2026

1. Introduction

The rapid advancement of artificial intelligence (AI) technologies has created unprecedented opportunities and challenges for educational systems worldwide. As AI tools become increasingly sophisticated and accessible, educators find themselves at the forefront of a technological transformation that demands new forms of literacy and pedagogical expertise (Holmes et al., 2019; Zawacki-Richter et al., 2019). In response, recent policy initiatives, including UNESCO's Beijing Consensus on AI in Education (2019) and the European Commission's Digital Education Action Plan (2021), have identified teacher preparation as a critical factor for the successful and equitable integration of AI in education.

However, the integration of AI in education extends far beyond mere technical proficiency. Teachers must develop comprehensive AI literacy encompassing an understanding of underlying mechanisms, critical evaluation capabilities, and ethical reasoning skills (Long & Magerko, 2020; Walter, 2024). Recent research further emphasizes that true AI literacy requires not only technical understanding but also prompt engineering proficiency and enhanced critical thinking skills to navigate AI's transformative impact on educational settings (Walter, 2024).

Despite a growing recognition of this need, empirical research on how teachers actually develop such multifaceted competencies remains remarkably limited. A recent scoping review starkly concluded that AI literacy is "a globally emerging research topic in education but almost absent in the context of teacher education" (Casal-Otero et al., 2023, p. 1). This study directly addresses this critical gap by investigating the *process* through which future teachers develop AI literacy across multiple, interdependent dimensions. Unlike previous research that has focused primarily on learning outcomes or program descriptions, we aim to illuminate the nature of the learning journey itself. By examining teachers' evolving reflections, we engage with broader questions that motivate this work: whether AI literacy development involves conceptual restructuring rather than additive accumulation; how teachers adapt pedagogical thinking across distinct AI-literacy dimensions; and what affective and metaphorical patterns characterize this journey. We address these broader questions indirectly, through a set of empirical research questions (presented in §2.4) that operate at the level of observable linguistic and thematic patterns in reflective writing. Interpretive claims about transformative learning are advanced in dialogue with – rather than as direct readings of – these empirical patterns. Methodologically, the study integrates computational text analysis with qualitative thematic analysis applied to the same corpus. We argue (§3.1) that the two strands generate analytical value not through redundancy but through complementarity at different levels of representation of reflective writing. Understanding this developmental process is not a purely theoretical concern: it bears directly on how teacher-education programs sequence content, design reflective tools, and pace the cognitive and affective transitions they ask preservice teachers to undertake.

2. Theoretical Framework

2.1 AI literacy as multidimensional competence

We frame AI literacy as a multidimensional competence that integrates cognitive understanding of concepts and models, operational capacity to work with data and tools, critical awareness of bias and societal implications, and ethical judgment oriented to professional responsibility. Rather than a purely technical skillset, AI literacy in teacher education entails developing shared language and categories that support responsible decision-making in context. This perspective aligns with accounts that emphasize competencies and design considerations for AI literacy and with policy documents that call for ethically grounded adoption in education (Long & Magerko, 2020; Laupichler et al., 2022; Wang, Rau, & Yuan, 2023; Grover & Pea, 2013; UNESCO, 2019; European Commission, 2021).

2.2 Transformative learning in technology integration

Transformative learning focuses on shifts in frames of reference through experience, reflection, and dialogue. Applied to technology integration, the central issue is not only what teachers learn about AI tools but how they reframe assumptions to reason about opportunities, constraints, and classroom consequences. Our analyses operationalize this stance by examining linguistic signals of discontinuity across phases together with markers of within-person consolidation – patterns consistent with reframing rather than additive accumulation (Mezirow, 1991; Taylor, 2008; Schön, 1983).

2.3 Metacognitive reflection as mechanism

Metacognitive reflection enables learners to monitor understanding, evaluate implications, and plan action. Brief, structured diaries collected at regular intervals provide a practical lens on these processes, capturing stance, uncertainty, and evolving categories. In teacher education, such journaling is a well-established practice to support professional growth and to surface reasoning that is otherwise tacit (Moon, 2006; Lipman, 2003).

2.4 Metacognitive reflection as mechanism

We investigate:

- **RQ1**: How do cohort-level lexical and topical patterns evolve across the four modules?
- **RQ2**: Do individual reflections show increasing coherence over time?
- **RQ3**: How do sentiment and metaphorical framings change alongside content?
- **RQ4**: How do qualitative themes corroborate and interpret computational trends?

A note on levels of analysis. The constructs framing this study AI literacy as multidimensional competence, transformative learning, metacognitive reflection are theoretical. The research questions above operate at a more proximal level: observable patterns in reflective writing (lexical, topical, affective, thematic). We do not claim that linguistic indicators directly measure transformative learning. Rather, following established uses of computational text analysis in education research, we treat such indicators as warrants for interpretive claims, jointly with qualitative thematic analysis. The integration of the two strands (§3.1) is what enables the move from descriptive patterns to interpretive readings, and it is also what distinguishes the present study from work that relies on either approach alone.

3. Method

3.1 Study design

We employed a convergent mixed-methods design (Creswell & Plano

Clark, 2011) in which computational analyses and qualitative thematic coding were conducted in parallel on the same corpus and then integrated. The two strands operate at different levels of representation of the data, which is what motivates their integration rather than treating them as parallel interpretations. *Computational analyses* (TF-IDF similarity, lexical diversity, topic modeling, sentiment) process the entire corpus and produce surface-level, cohort-wide indicators that no human coder could observe at this scale; they answer questions of the form "what patterns recur across all 249 diaries?" They are deliberately blind to the semantics that participants themselves attach to their reflections. *Qualitative thematic analysis*, conducted by two independent coders on the full corpus, works at the level of meaning, narrative logic, and participants' own voice – dimensions that TF-IDF representations flatten by construction. It answers questions of the form "how do participants articulate what is changing, and in what terms?"

The integration generates analytical value in three concrete ways. *First*, the computational layer extends the qualitative findings beyond the cases the coders inspected most closely: when thematic analysis identifies a transition from fear-based to partnership metaphors, similarity and topic-modeling results indicate whether that transition is a localized observation or a corpus-wide pattern. *Second*, the qualitative layer authorizes interpretive readings of the computational indicators: cosine similarity of ~0.08 between reflection phases is, on its own, a statistical fact about lexical overlap; only the thematic analysis allows us to read this pattern as *conceptual* restructuring rather than mere vocabulary turnover. *Third*, the procedural coupling worked iteratively: extreme cases identified computationally (diaries at the tails of similarity, coherence, or sentiment distributions) were re-examined qualitatively, and recurrent themes from coding informed how we labeled and interpreted LDA topics. We do not claim methodological independence between the two strands they share corpus and framework but rather methodological *complementarity*: each approach compensates for a specific limitation of the other.

3.2 Study design

The study involved 63 Italian future teachers enrolled in a university-based professional development program. Participants represented primary, lower-secondary, and upper-secondary contexts and subject areas, with heterogeneous prior exposure to AI. Table 1 summarizes demographics and teaching backgrounds.

Table 1 - Participant Demographics (N = 63)

Characteristic	***N***	***%***
Female	56	89%
Teaching Experience >15 years	32	51%
STEM Subjects	21	34%
No Prior AI Experience	41	65%

3.3 Intervention: four-module curriculum

The 16-week program comprised four sequential modules aligned with the AI-literacy dimensions. Module 1 (Cognitive) addressed key concepts, terminology, and representations; Module 2 (Operational) emphasized algorithms, data, and practical activities; Module 3 (Critical) explored bias, accountability, and broader societal issues; Module 4 (Ethical) focused on professional responsibilities and policy. Each module included readings, cases, and collaborative tasks.

3.4 Data collection

After each module, participants produced brief reflective diaries structured in three sections corresponding to the three dimensions targeted by the study: (i) *Initial Beliefs* - what the participant believed about the module's topic; (ii) *Current Knowledge* - what the participant had learned during the module; and (iii) *Pedagogical Applications* - how the participant envisaged transferring the module's content into teaching practice, with reference to one or more concrete activities and practical examples. The three prompts were open-ended questions held constant across the four modules; only the topic referent (cognitive, operational, critical, ethical) varied. Each section had a soft length indication of 1,000 characters, which corresponds to approximately 150-180 words and was sufficient to allow participants to develop a coherent reflection without compression. The diary was delivered through an online Google Form embedded within the program's Moodle environment, so that participants could readily access it alongside the module materials. The diary also included an item in which participants selected an image representing their envisaged didactic scenario and described it in writing; the descriptive text was included in the corpus on the same basis as the other sections, while the image selection itself was not analyzed as a separate visual datum. The full text of the prompts, in the original Italian and in English translation, is reported in Appendix B. Diaries were de-identified by replacing names with participant codes (P01-P63) prior to any analysis. Table 2 reports corpus characteristics by module.

Table 2 - Corpus Characteristics by Module

Module	*Dimension*	*N*	*Total Words*	*TTR*
1	Cognitive	63	25,092	0.2139
2	Operational	62	24,277	0.2033
3	Critical	61	27,026	0.1968
4	Ethical	63	26,560	0.1926

3.5 Data analysis

3.5.1 Qualitative analysis

We conducted thematic analysis with an iterative codebook anchored to the theoretical framework. Two coders independently labelled a stratified subset of diaries, discussed disagreements, refined code definitions, and then coded the full corpus. We documented decisions in an audit trail, reported representative excerpts for each theme, and quantified reliability using Cohen's κ with interpretation following established guidelines (Braun & Clarke, 2006; Landis & Koch, 1977).

Table 3 - Inter-rater Reliability Scores

Code Category	*Cohen's κ*	*Interpretation*
Attitude Positive	0.859	Almost Perfect
Attitude Negative	0.848	Almost Perfect
Conceptual Knowledge	0.859	Almost Perfect
Operational Knowledge	0.792	Substantial
Critical Knowledge	0.860	Almost Perfect
Ethical Knowledge	0.745	Substantial
Application Coding	0.872	Almost Perfect
Application Collaborative	0.867	Almost Perfect

3.5.2 Computational analysis

Preprocessing. Italian texts were processed with spaCy (it_core_news_sm), lowercased and lemmatized; punctuation, numerals, and non-alphabetic tokens were removed. We used the default Italian stop list and extended it with a curated list of high-frequency function lemmas (e.g., auxiliaries, prepositions). Pre-processed outputs were stored as lemma lists per diary section (Beliefs, Knowledge, Applications), ensuring a consistent basis for all downstream analyses (Manning, Raghavan, & Schütze, 2008).

Representations and similarity. Lemma lists were concatenated and vectorized with TF-IDF (unigrams). For cohort-level inter-diary similarity, we aggregated diaries by module and section and computed cosine similarity between sections across phases (Beliefs↔Knowledge; Knowledge↔Applications). As a robustness check, we also computed Jaccard similarity on unique lemma sets. To estimate within-person consolidation, we computed intra-diary coherence by comparing each participant's sections across consecutive modules in the same TF-IDF space (Salton & Buckley, 1988).

A consideration on interpretation. Because the diary structure prompts participants to address three distinct themes (beliefs, knowledge, applications), a degree of lexical separation between sections is expected on design grounds: discussing initial beliefs naturally invokes a different content vocabulary than discussing learned concepts or pedagogical activities. We therefore interpret inter-section cosine similarity not as a free indicator of conceptual distance, but as a measure of how strongly the three thematic sections lexically diverge, against the baseline expectation that some divergence is built in. The interpretive weight of this metric in §4.1 rests on its *configuration* with other indicators within-participant coherence, lexical diversity, sentiment trajectory and on its convergence with the qualitative analysis in §4.3.

Lexical diversity. We calculated type-token ratio (TTR) per section and module as the number of unique lemmas divided by tokens.

Topics and sentiment. We trained a Latent Dirichlet Allocation model with k = 5 topics and examined shifts in topic prevalence across modules; sentiment at section level was obtained with a pretrained Italian pipeline and compared across modules using non-parametric tests (Blei, Ng, & Jordan, 2003; Röder, Both, & Hinneburg, 2015; Devlin et al., 2019). All analyses used fixed random seeds where applicable and operated on the same pre-processed representations to ensure comparability.

Topic number selection and justification. We inspected topic coherence (c_v) across k = 3-10 and qualitative interpretability of top-words. Coherence peaked/plateaued around k = 5-6; k = 5 offered the best balance between parsimony and thematic separation and aligned with our four analytic dimensions while allowing one cross-cutting theme (Röder et al., 2015). To ensure stability, we verified near-identical top-word sets over multiple random starts; perplexity was not used as a primary criterion due to its weaker correspondence with human interpretability.

4. Results

For readability, when reporting top-terms we provide English glosses of the original Italian lemmas; the model and all preprocessing operated entirely on Italian texts.

4.1 Inter-Section Lexical Discontinuity and Within-Person Coherence

Descriptive level. As shown in Table 4, inter-section cosine similarities between participants' initial beliefs, acquired knowledge, and proposed pedagogical applications were consistently low across all modules, with mean values between 0.051 and 0.095. These values indicate substantial lexical discontinuity between the three sections of reflection: the vocabulary used to describe initial beliefs shows minimal overlap with the lexicon used to articulate acquired knowledge and to design teaching activities. At the same time, intra-diary coherence each participant's lexical consistency across consecutive modules remained high throughout the program, with mean values around 0.80. The two patterns coexist: at the cohort level, the lexicon associated with beliefs, knowledge, and applications becomes increasingly distinct; at the individual level, each participant's vocabulary remains internally consistent.

Interpretive level. This combined pattern is consistent with conceptual restructuring rather than additive accumulation. As noted in §3.5.2, the tripartite thematic structure of the diary contributes a baseline component to inter-section separation; the relevant pattern is therefore not the low absolute similarity, but its *configuration* with stable within-participant coherence. If learners were merely appending new information to existing frameworks while remaining lexically consistent within each diary, we would expect substantial overlap between the lexicon of "initial beliefs" and that of "acquired knowledge" *within those same coherent diaries* yet this is precisely the pattern we do not observe. We treat this as an indicator pattern, not as direct evidence of transformation; the qualitative analysis in §4.3 provides an independent reading of the same corpus, and we return to the integrated interpretation in §5.

Table 4 - Similarity Analysis: Evidence for Transformative Learning

Module	***Beliefs-Knowledge***	***Knowledge-Applications***	***Intra-diary Coherence***
Cognitive	0.079 ± 0.045	0.067 ± 0.038	0.801 ± 0.156
Operational	0.090 ± 0.052	0.066 ± 0.041	0.764 ± 0.142
Critical	0.089 ± 0.048	0.095 ± 0.054	0.823 ± 0.167
Ethical	0.081 ± 0.043	0.075 ± 0.047	0.795 ± 0.149

Note: Values represent mean ± standard deviation

4.2 Lexical Evolution and Specialization

The development of a more sophisticated understanding was further corroborated by lexical analysis. The Type-Token Ratio (TTR) showed a progressive decline across the four modules, from 0.2139 in the Cognitive module to 0.1926 in the Ethical module. This decrease in lexical diversity indicates that participants were gradually converging on a more specialized and precise vocabulary required for a professional discourse on AI, moving away from generic language.

TF-IDF analysis confirmed that this specialization was dimension-specific, with participants acquiring the relevant terminology for each module (Table 5). For instance, 'intelligenza' (intelligence) and 'umano' (human) were highly significant in the Cognitive module, while 'algoritmo' (algorithm) and 'dati' (data) became central in the Operational one. Subsequently, terms like 'bias' and 'critico' (critical) dominated the Critical module, and 'etica' (ethics) and 'responsabilità' (responsibility) became the key terms in the final Ethical dimension. This demonstrates a structured acquisition of dimension-specific language.

Table 5 - Top Distinctive Terms by Module (TF-IDF Analysis)

Rank	***Cognitive***	***Operational***	***Critical***	***Ethical***
1	intelligence (6.36)	algorithm (8.92)	bias (9.18)	ethics (10.24)
2	human (5.11)	data (7.45)	critical (8.23)	responsibility (8.67)
3	artificial (4.83)	instructions (6.12)	thinking (7.89)	decisions (7.23)
4	machine (4.37)	computational (5.67)	prejudices (5.45)	privacy (6.45)
5	definition (4.21)	machines (4.98)	images (4.87)	dimension (5.89)

4.3 Thematic Analysis: A Transformative Learning Journey

Whereas the computational results in §4.1-4.2 describe patterns at the lexical surface, the qualitative analysis allows us to interpret those patterns by examining how participants themselves articulated their changing understanding. The qualitative analysis of the diary entries revealed a clear and systematic progression that aligns remarkably well with Mezirow's phases of transformative learning, providing a narrative context for the quantitative findings.

- **Phase 1: Disorienting Dilemma.** Initial diary entries were characterized by metaphors of profound uncertainty and apprehension. Participants described AI as an "ocean," a "secret garden," a "labyrinth," or even a "runaway horse," expressing a sense of being overwhelmed by its implications for their professional identity. One teacher wrote: *"AI is like standing at the edge of an unknown ocean – vast, deep, and somewhat frightening"* (P23, Module 1).
- **Phase 2: Critical Self-Examination.** As the course progressed, participants began to critically examine their initial fears and assumptions. They

increasingly recognized the limitations of their existing mental frameworks for understanding technology. A key reflection from this phase was: *"I realize my resistance comes from fear of becoming obsolete... but maybe that's not the real issue"* (P41, Module 2).

- **Phase 3: Exploration and Building Competence.** The operational and critical modules were pivotal for demystifying AI. Through hands-on activities and critical debate, participants built a more grounded and technical understanding, moving from abstract fears to concrete knowledge. This shift was evident in statements like: *"Now I see it's not magic, it's mathematics – patterns in data, not mysterious intelligence"* (P55, Module 2).
- **Phase 4: Integration and Transformation.** The final reflections in the ethical module demonstrated a fully integrated and transformed understanding. AI was no longer seen as an external, threatening force, but as a potential collaborator in the pedagogical process. A powerful concluding reflection captured this new perspective: *"AI has become my teaching partner – I provide human understanding, ethical guidance, and emotional support; it provides computational power and pattern recognition"* (P19, Module 4).

4.4 Pedagogical Adaptation Across Dimensions

The participants not only transformed their understanding of AI but also adapted their pedagogical thinking. A Chi-square analysis revealed statistically significant differences in their preferences for pedagogical scenarios across the four modules ($\chi^2(15) = 58.928$, $p < 0.001$, Cramér's $V = 0.278$), indicating a systematic adaptation of teaching strategies.

As shown in Table 6, for the Cognitive dimension, individual reflection was the most frequently proposed activity (32.3%). This shifted towards collaborative laboratory work for the Operational dimension (32.3%). Most notably, there was a strong move toward group discussion for the Critical (31.7%) and Ethical (41.5%) dimensions, suggesting a recognition that these competencies require social negotiation of values. From a curriculum-design perspective, this finding suggests that pedagogical scenarios for AI-literacy programs should be aligned to the dimension being addressed: technical content benefits from individual and laboratory work, while critical and ethical content calls for discursive formats in which values can be socially negotiated rather than transmitted.

Table 6 - Pedagogical Scenario Preferences by AI Literacy Dimension

Scenario	Cognitive	Operational	Critical	Ethical
Individual Reflection	21 (32.3%)	10 (16.1%)	13 (20.6%)	6 (9.2%)
Collaborative Laboratory	14 (21.5%)	20 (32.3%)	9 (14.3%)	18 (27.7%)
Group Discussion	8 (12.3%)	6 (9.7%)	20 (31.7%)	27 (41.5%)
Individual + Technology	12 (18.5%)	13 (21.0%)	2 (3.2%)	2 (3.1%)
Analysis of Artifacts	6 (9.2%)	4 (6.5%)	11 (17.5%)	10 (15.4%)
Formal/Scientific	4 (6.2%)	9 (14.5%)	8 (12.7%)	2 (3.1%)

4.5 Metaphorical and Emotional Evolution

Metaphors evolved from AI as an opaque "black box" or "wave" to AI as a "toolbox," "co-teacher," or "laboratory," signalling reframed agency and professional positioning. This journey from fear to partnership was echoed in the sentiment analysis results (Table 7), which show a nuanced emotional pattern. The initial enthusiasm for applying AI became progressively tempered by a growing awareness of critical and ethical complexities, indicating a maturation from naive optimism to a more responsible and critical stance.

Table 7 - Sentiment Analysis Results

Section	Module 1	Module 2	Module 3	Module 4
Initial Beliefs	0.152 ± 0.234	0.143 ± 0.267	0.149 ± 0.245	0.141 ± 0.256
Current Knowledge	0.298 ± 0.189	0.287 ± 0.203	0.291 ± 0.197	0.279 ± 0.211
Applications	0.476 ± 0.167	0.389 ± 0.178	0.312 ± 0.189	0.241 ± 0.201

5. Discussion

5.1 Theoretical implications

Our findings align with transformative learning accounts: linguistic discontinuities between phases co-occur with within-person consolidation, suggesting reframing rather than mere accretion. The multidimensional AI-literacy model offers a useful scaffold to interpret how beliefs, knowledge, and applications co-evolve. This study offers one of the first empirically grounded accounts of transformative learning in the context of AI literacy development for teachers. The consistent pattern of low inter-diary similarity combined with high intra-diary coherence provides quantitative indicators consistent with conceptual restructuring rather than additive accumulation – an interpretation that is reinforced, not established, by the converging qualitative evidence. The clear progression from fear-based metaphors to partnership conceptualizations provides empirical support for theoretical models that propose staged development in AI literacy (Yim, 2024).

The convergence between computational and qualitative findings reported in §4 should be read in light of the design constraints discussed in §3.1: the two strands operate on the same corpus and within the same theoretical framework, and the qualitative analysis informed our interpretation of the computational indicators. Their convergence is therefore not independent corroboration in the strong sense, but a coherent reading of the corpus from two different levels of representation. We see this as a feature rather than a limitation: the computational layer provides scale and surface-level pattern detection that no qualitative effort could match on 249 diaries, while the qualitative layer provides the interpretive grounding without which the computational signals would remain descriptive statistics.

5.2 Pratical implications

Teacher-education curricula can cultivate sustainable AI literacy by sequencing experiences that (a) name core concepts, (b) surface biases and societal stakes, (c) connect ideas to pedagogical routines, and (d) foreground ethical reasoning. Short, structured reflective prompts – analysed responsibly – provide timely feedback on cohort trajectories and inform instructional adjustments. The 16-week timeframe of this program was a key factor in allowing for genuine conceptual transformation; shorter programs are likely to result in only surface-level changes (Dilek et al., 2025). Finally, acknowledging and creating space for the emotional journey of teachers is crucial, supporting them as they move from initial anxiety towards a confident and critical partnership with AI (Lee et al., 2024). More concretely, the pattern of results suggests four design heuristics: (i) sequencing dimensions from cognitive scaffolding to ethical reasoning rather than front-loading technical content; (ii) using short, structured reflective prompts not only as assessment but as visible scaffolds that legitimize the affective dimension of the transition; (iii) varying pedagogical formats by dimension, as the Chi-square pattern in §4.4 indicates; and (iv) protecting sufficient program length to allow conceptual reorganization to occur, rather than compressing AI-literacy training into one-off workshops.

5.3 Limitations

Analyses rely on brief written reflections and NLP measures that, while informative, only approximate underlying cognition and emotion. Topic and sentiment models have known constraints in Italian, and TTR is sensitive to text length. Findings are situated in one program and context; generalization requires caution. The convergent design pairs two analytic strands that share a corpus and a theoretical framework; their integration provides complementary rather than independent evidence, as discussed in §5.1. The reflective diary is itself a structuring instrument: by inviting participants to address three distinct dimensions (beliefs, knowledge, applications), it imposes a thematic partition that contributes some baseline component to the inter-section lexical separation reported in §4.1. We have addressed this point methodologically (§3.5.2) and interpretively (§4.1, §5.1) by grounding our claims on configurations of indicators rather than on inter-section similarity in isolation, and by integrating the computational layer with qualitative thematic analysis. A study design comparing diaries with partitioned versus non-partitioned prompts would provide stronger isolation of participant-driven from instrument-driven variance, and we see this as a productive direction for replication work. The absence of a control group means we cannot definitively attribute the observed transformation solely to our program's specific features.

5.4 Future research directions

Future work should combine diary analytics with classroom artifacts and observations, examine longer-term retention and transfer, and compare alternative curricular sequences. Model-transparent sentiment and topic models fine-tuned on educational Italian corpora would improve measurement validity. Comparative designs that test transformative approaches against more traditional training could also help isolate the most effective components of professional development (Ding et al., 2024).

6. Conclusion

This study provides the first mixed-methods empirical demonstration of transformative learning in AI literacy development, contributing both theoretical insights for the field and a practical, scalable methodology for future research. Through an integrated analysis of 249 reflective diaries, we documented how future teachers fundamentally restructure their understanding across the cognitive, operational, critical, and ethical dimensions of AI. The documented journey from viewing AI as a threatening “unknown territory” to embracing it as an “ethical partnership” represents not merely the acquisition of new knowledge, but a profound reconstruction of professional identity. For teacher educators and curriculum designers, this implies that effective AI-literacy programs cannot be reduced to technical training: they must accommodate, and explicitly scaffold, the conceptual and affective transitions through which professional identity is reorganized.

As artificial intelligence continues to reshape the educational landscape, supporting teachers through this challenging but essential transformative journey is paramount. It is the key to realizing AI's vast educational potential while simultaneously preserving human values and pedagogical wisdom. The implications of our work extend beyond AI literacy to the broader question of how we, as an educational community, can better support educators in adapting to an era of relentless technological change. Ultimately, the challenge is not merely to teach teachers *about* AI, but to facilitate their transformation into reflective, critical, and empowered professionals who can confidently navigate, critique, and shape AI's role in the future of education.

Declarations

Ethics approval and consent to participate

Ethics approval. The study involved voluntary, non-interventional activities (anonymous reflective diaries) with adult university students and collected no sensitive personal data as defined under Article 9 of the EU General Data Protection Regulation (GDPR). According to the policy of the *Comitato Etico per la Ricerca* at the University of Florence, non-clinical research that does not involve sensitive personal data and uses anonymized datasets does not require formal submission to, or approval by, the Committee. Therefore, formal ethics approval was not required for this research. The study was conducted in accordance with the University of Florence Code of Ethics and applicable data protection regulations (EU GDPR 2016/679 and Italian Legislative Decree 101/2018).

Consent to participate. All participants received an information sheet describing the purpose of the research, procedures, and data handling. Participation was voluntary, and by completing the activities participants provided informed consent. No identifiable information was collected.

Consent for publication

Participants consented to the publication of anonymized, aggregated results and illustrative, non-identifiable excerpts from diary entries. No identifying information will be disclosed.

Availability of data and materials

The data supporting this study are not openly available because the raw diary entries contain information that could compromise participant privacy. Raw entries will remain confidential. Anonymized analysis scripts, coding schemes, and other relevant materials will be made available by the corresponding author upon reasonable request after publication.

Clinical trial number: not applicable

Competing interests
The authors declare that they have no competing interests.

Funding statement
This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

References

Akgun S., & Greenhow C. (2022). Artificial intelligence in education: Addressing ethical challenges in K-12 settings. *AI and Ethics*, 2(3): 431-440.

Ayanwale M. A., Adelana O. P., Molefi R. R., Adeeko O., & Ishola A. M. (2024). Examining artificial intelligence literacy among pre-service teachers for future classrooms. *Computers and Education Open*, 6, 100179.

Blei D. M., Ng A. Y., & Jordan M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3: 993-1022.

Braun V., & Clarke V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, 3(2): 77-101.

Brooks R. E., & Brooks A. K. (2024). An emerging transformative learning journey to foster sustainability leadership in professional development programs. *Journal of Transformative Education*. Doi: 10.1177/15413446241255909.

Casal-Otero L., Catala A., Fernández-Morante C., Taboada M., Cebreiro B., & Barro S. (2023). AI literacy in K-12: A systematic literature review. *International Journal of STEM Education*, 10(1), 29.

Creswell J. W., & Plano Clark V. L. (2011). *Designing and conducting mixed methods research* (2nd ed.). Sage Publications.

Devlin J., Chang M. W., Lee K., & Toutanova K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT* (pp. 4171-4186).

Dilek M., Baran E., & Aleman E. (2025). AI literacy in teacher education: Empowering educators through critical co-discovery. *Journal of Teacher Education*, 76(1). Doi: 10.1177/00224871251325083.

Ding A. C. E., Shi L., Yang H., & Choi I. (2024). Enhancing teacher AI literacy and integration through different types of cases in teacher professional development. *Computers and Education Open*, 6, 100178.

Du H., Sun Y., Jiang H., Muhammad A., & Liu J. (2024). Exploring the effects of AI literacy in teacher learning: An empirical study. *Humanities and Social Sciences Communications*, 11, 559.

Ertmer P. A., & Ottenbreit-Leftwich A. T. (2010). Teacher technology change: How knowledge, confidence, beliefs, and culture intersect. *Journal of Research on Technology in Education*, 42(3), 255-284.

European Commission (2021). *Digital Education Action Plan 2021-2027*. Publications Office of the European Union.

Grover S., & Pea R. (2013). Computational thinking in K-12: A review of the state of the field. *Educational Researcher*, 42(1): 38-43.

Holmes W., Bialik M., & Fadel C. (2019). *Artificial intelligence in education: Promises and implications for teaching and learning*. Center for Curriculum Redesign.

Holmes W., Porayska-Pomsta K., Holstein K., Sutherland E., Baker T., Shum S. B., ... & Koedinger K. R. (2022). Ethics of AI in education: Towards a community-wide framework. *International Journal of Artificial Intelligence in Education*, 32(3): 504-526.

Kim C., Kim M. K., Lee C., Spector J. M., & DeMeester K. (2013). Teacher beliefs and technology integration. *Teaching and Teacher Education*, 29: 76-85.

Koehler M. J., Mishra P., Kereluik K., Shin T. S., & Graham C. R. (2014). The technological pedagogical content knowledge framework. In: *Handbook of research on educational communications and technology* (pp. 101-111). Springer.

Landis J. R., & Koch G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33(1): 159-174.

Laupichler M. C., Aster A., Schirch J., & Raupach T. (2022). Artificial intelligence literacy in higher and adult education: A scoping literature review. *Computers and Education: Artificial Intelligence*, 3, 100101.

Lee D., Arnold M., Srivastava A., Plastow K., Strelan P., Ploeckl F., Lekkas D., & Palmer E. (2024). The impact of generative AI on higher education learning and teaching: A study of educators' perspectives. *Computers and Education: Artificial Intelligence*, 6, 100221.

Lipman M. (2003). *Thinking in education* (2nd ed.). Cambridge University Press.

Long D., & Magerko B. (2020). What is AI literacy? Competencies and design considerations. In: *Proceedings of CHI 2020* (pp. 1-13).

Manning C. D., Raghavan P., & Schütze H. (2008). *Introduction to information retrieval.* Cambridge University Press.

McConnell T. J., Parker J. M., & Eberhardt J. (2019). Problem-based learning for responsive and transformative teacher professional development. *Global Journal of Transformative Education*, 1(1), 18-25.

Mezirow J. (1991). *Transformative dimensions of adult learning*. Jossey-Bass.

Mishra P., & Koehler M. J. (2006). Technological pedagogical content knowledge: A framework for teacher knowledge. *Teachers College Record*, 108(6): 1017-1054.

Moon J. A. (2006). *Learning journals: A handbook for reflective practice and professional development* (2nd ed.). Routledge.

Ng D. T. K., Leung J. K. L., Chu K. W. S., & Qiao M. S. (2021). AI literacy: Definition, teaching, evaluation and ethical issues. In: *Proceedings of ITiCSE 2021* (pp. 504-509).
Röder M., Both A., & Hinneburg A. (2015). Exploring the space of topic coherence measures. In: *Proceedings of WSDM* (pp. 399-408).
Russell S., & Norvig P. (2020). *Artificial intelligence: A modern approach* (4th ed.). Pearson.
Salton G., & Buckley C. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing & Management*, 24(5): 513-523.
Schön D. A. (1983). *The reflective practitioner: How professionals think in action*. Basic Books.
Su J., Zhong Y., & Ng D. T. K. (2022). A meta-review of literature on educational approaches for teaching AI at the K-12 levels in the Asia-Pacific region. *Computers and Education: Artificial Intelligence*, 3, 100065.
Tan Q., & Tang X. (2025). Unveiling AI literacy in K-12 education: A systematic literature review of empirical research. *Interactive Learning Environments*. Doi: 10.1080/10494820.2025.2482586.
Taylor E. W. (2008). Transformative learning theory. New Directions for Adult and *Continuing Education*, (119): 5-15.
Touretzky D., Gardner-McCune C., Martin F., & Seehorn D. (2019). Envisioning AI for K-12: What should every child know about AI? In: *Proceedings of AAAI* (Vol. 33, pp. 9795-9799).
Tuomi I. (2018). *The impact of artificial intelligence on learning, teaching, and education*. Publications Office of the European Union.
UNESCO (2019). *Beijing consensus on artificial intelligence and education*. United Nations Educational, Scientific and Cultural Organization.
Walter Y. (2024). Embracing the future of Artificial Intelligence in the classroom: The relevance of AI literacy, prompt engineering, and critical thinking in modern education. *International Journal of Educational Technology in Higher Education*, 21, 15.
Wang B., Rau P. L. P., & Yuan T. (2023). Measuring user competence in using artificial intelligence: Validity and reliability of artificial intelligence literacy scale. *Behaviour & Information Technology*, 42(9): 1324-1337.
Wing J. M. (2006). Computational thinking. *Communications of the ACM*, 49(3): 33-35.
Yim I. H. Y. (2024). A critical review of teaching and learning artificial intelligence (AI) literacy: Developing an intelligence-based AI literacy framework for primary school education. *Computers and Education Open*, 5, 100162.
Yim I. H. Y., & Wegerif . (2024). Teachers' perceptions, attitudes, and acceptance of artificial intelligence (AI) educational learning tools: An exploratory study on AI literacy for young students. *Future in Educational Research*, 2(1), e65.
Zawacki-Richter O., Marín V. I., Bond M., & Gouverneur F. (2019). Systematic review of research on artificial intelligence applications in higher education. International *Journal of Educational Technology in Higher Education*, 16(1): 1-27.
Zipf G. K. (1949). *Human behavior and the principle of least effort*. Addison-Wesley.

Appendix A. Topic coherence (c_v) selection for LDA

```
# Appendix A: Compute c_v coherence for k in [3..10]
# Requires: gensim
import random
import gensim
from gensim.corpora import Dictionary
from gensim.models import CoherenceModel

# docs_lemmatized: list of lists of lemmas (one diary per list)
# Example: docs_lemmatized = [['ai','modello','dato',...], ...]

random.seed(100)

def compute_cv_coherence(docs_lemmatized, k_values=range(3,11)):
    dictionary = Dictionary(docs_lemmatized)
    corpus = [dictionary.doc2bow(doc) for doc in docs_lemmatized]
    scores = []
    for k in k_values:
        lda = gensim.models.LdaModel(
            corpus=corpus,
            id2word=dictionary,
            num_topics=k,
            random_state=100,
            passes=10,
            alpha='auto'
        )
        cm = CoherenceModel(model=lda, texts=docs_lemmatized,
dictionary=dictionary, coherence='c_v')
        scores.append(cm.get_coherence())
    return list(k_values), scores
```

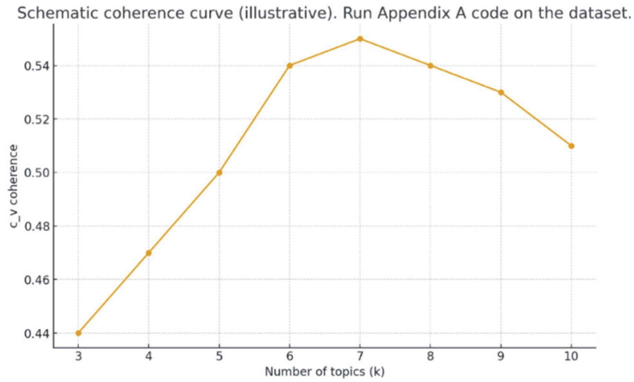

Figure A1 - Schematic c_v coherence across k = 3-10. Note: illustrative only; regenerate by running the Appendix A code on the actual corpus

Appendix B. Reflective diary prompts

Participants completed a reflective diary after each of the four modules. The diary was delivered through an online Google Form embedded within the program's Moodle environment, so that participants could readily access it alongside the module materials. The prompts were held constant across the four modules; only the topic referent (the specific dimension of AI literacy addressed in that module) varied. The first three items had a soft length indication of 1,000 characters per section (~150-180 words).

Italian (original)

1. *Quali erano le tue credenze sull'argomento trattato a lezione?* (1000 caratteri)
2. *Che cosa sai ora dell'argomento trattato?* (1000 caratteri)
3. *Come userai le conoscenze apprese in questo percorso? Pensa ad una o più attività concrete e illustrale con esempi pratici.* (1000 caratteri)
4. *Sulla base delle tue riflessioni, scegli l'immagine che maggiormente si avvicina allo scenario d'uso didattico a cui hai pensato.*
5. *Descrivi ora l'immagine scelta e illustra l'applicazione didattica a cui hai pensato.*

English translation

1. *What were your beliefs about the topic discussed in class?* (1,000 characters)
2. *What do you know now about the topic discussed?* (1,000 characters)
3. *How will you use the knowledge acquired in this program? Think of one or more concrete activities and illustrate them with practical examples.* (1,000 characters)
4. *Based on your reflections, choose the image that best matches the didactic-use scenario you envisioned.*
5. *Now describe the image you chose and illustrate the didactic application you had in mind.*